Critical Incident Report

Incident Management

PROBLEM TICKET # - TBD	**INCIDENT TICKET # -** INC000000086503/86633
DATE/TIME ISSUE STARTED - TBD	**BUSINESS UNIT/SYSTEMS AFFECTED – CDF; D0; CMS; FermiGrid; FermiCloud (test); FNPC**
DATE/TIME ISSUE REPORTED – 3:13P on Tuesday, June 7.	**GROUP ASSIGNED – Facilities Operations**
DATE/TIME ISSUE RESOLVED/ RESTORED – Partial restoration at 5:25P June 7. Full restoration at 6:10A on June 9.	**CRITICAL INCIDENT MANAGER – Brian McKittrick**
OUTAGE DURATION – At Critical Priority: 2 hours. At High Priorty: 32.5 Hours at High Priority. Total duration of ~34.5 hours.	**INCIDENT RESPONSE INFORMATION Room FCC1 Audio Bridge – 866-740-1260 code 1113472**

Send out email to Division Management to indicate a critical incident has been declared. Briefly describe the issue.

Update Service Desk.FNAL.gov; Tweet; Remedy Broadcast; SD-Broadcast Chat room where applicable

Determine of a Command Center is needed

Record Command Center participants below

Are all necessary stakeholders present? Do we need anyone else engaged?

Gain understanding of issue and current state (Log in Incident Description and Symptoms)

Identify Impact (Log in Business Impact Sections)

Establish time when incident began (Log in Incident Description and Symptoms)

Identify Ticket number

Identify which Customer(s) need to be communicated to

Intentionally Left Blank

Incident and Incident Review Participants (Roll Call)

Functional Area	Date	Name	Role
3:13P – Critical Incident Command Center	6/7		
ITSM	6/7	Brian McKittrick	Critical Incident Manager
ITSM	6/7	Jack Schmidt	Acting Service Manager
ITSM	6/7	Gerald Guglielmo	Problem Manager
FermiGrid/FermiCloud	6/7	Keith Chadwick	Technical
CMS	6/7	Jon Bakken	Technical
Scientific Computing	6/7	Stu Fuess	Technical
FEF	6/7	Jason Allen	Technical
REX	6/7	Margaret Votava	Technical

Scientific Computing	6/7	Qizhong Li	Liaison
Scientific Computing	6/7	Michael Diesburg	Liaison
Scientific Programs	6/7	Patty McBride	Management
Office of CIO	6/7	Mark Leininger	Management
Scientific Computing/ Facilities	6/7	Stephen Wolbers	Management
Facilities	6/7	Adam Walters	Technical
Facilities	6/7	Tim Kasza	Technical
Facilities	6/7	Jerry Bellandir	Technical
7:00A – Status with Facilities	6/8		
ITSM	6/8	Brian McKittrick	Critical Incident Manager
ITSM	6/8	Jack Schmidt	Service Level Manager
ITSM	6/8	Gerald Guglielmo	Problem Manager
ITSM	6/8	Tim Currie	Service Manager
Facilities	6/8	Adam Walters	Technical
Facilities	6/8	Tim Kasza	Technical
Facilities	6/8	Jerry Bellandir	Technical
Facilities	6/8	John Macnerland	Technical
8:00A – Status with Team	6/8		
ITSM	6/8	Brian McKittrick	Critical Incident Manager
ITSM	6/8	Jack Schmidt	Service Level Manager
ITSM	6/8	Gerald Guglielmo	Problem Manager
ITSM	6/8	Tim Currie	Service Manager
Facilities	6/8	Adam Walters	Technical
Facilities	6/8	Tim Kasza	Technical
Facilities	6/8	Jerry Bellandir	Technical
Facilities	6/8	John Macnerland	Technical
SNS	6/8	Ray Pasetes	Technical
Scientific Computing/ Facilities	6/8	Stephen Wolbers	Management
Scientific Programs	6/8	Patty McBride	Management
FEF	6/8	Jason Allen	Technical
Scientific Computing	6/8	Qizhong Li	Liaison
FermiGrid/FermiCloud	6/8	Keith Chadwick	Technical
CMS	6/8	Dave Fagan	Technical
3:00P – Status with team	6/8		
ITSM	6/8	Brian McKittrick	Critical Incident Manager
ITSM	6/8	Jack Schmidt	Service Level Manager
ITSM	6/8	Tim Currie	Service Manager
ITSM	6/8	Gerald Guglielmo	Problem Manager
CMS	6/8	Jon Bakken	Technical
CMS	6/8	Dave Fagan	Technical
FermiGrid/FermiCloud	6/8	Keith Chadwick	Technical
Facilities	6/8	Adam Walters	Technical
Scientific Computing/ Facilities	6/8	Stephen Wolbers	Management
Scientific Programs	6/8	Patty McBride	Management
FEF	6/8	Jason Allen	Technical
Scientific Computing	6/8	Su Fuess	Technical

Facilities	6/8	Tim Kasza	Technical
Core Services	6/8	Mark Kaletka	Technical
Facilities	6/8	Jerry Bellandir	Technical
Facilities	6/8	John Macnerland	Technical
6:00A Restoration Meeting	6/9		
ITSM	6/9	Brian Mckittrick	Critical Incident Manager
ITSM	6/9	Jack Schmidt	Service Level Manager
CMS	6/9	Jon Bakken	Technical
Facilities	6/9	Adam Walters	Technical
Facilities	6/9	Tim Kasza	Technical

Incident Summary and Details

GCC B CRAC unit 15 began having issues on Monday, June 6. These issues caused the other cooling units to struggle and temperatures reached a critical level.

Efforts were made by facilities to restore an acceptable temperature level, however at ~3:10 7 racks were shut down.

At ~3:30P all computing in GCC B was powered off.

Impact in Down State

All Computing in GCC B was unavailable from ~3:30P until ~5:25P.

FermiGrid - The ensemble of FermiGrid (cdf, cms, d0, gp) worker nodes were running ~18,000 slots immediately prior to the incident, following the incident, a total of~1800 slots were running.

FermiCloud – Will be left down for the duration of the cooling issues.

CMS – Impact on CMS captured in supporting pictures.

D0 - Estimating 5000 job lost when power was lost. Painful for users to search logs and resubmit jobs.

CDF - Approximately 4500 jobs lost when the power when out. Painful for users to search logs and resubmit jobs.

GPCF – completely down following the emergency shutdown.

Impact in degraded state

Service was restored to an acceptable level at ~5:25 on June 7. Service was restored to normal operation at 6:10A on Thursday, June 9.

FermiCloud – Will be left down for the duration of the cooling issues.

D0 - capacity has dropped from 6500 slots to 5200 slots. There are 2 batch system for D0 currently impacted. One is running at 75% and the other is running at 90%. Production could fall behind possibly impacting conferences and deadlines

CDF – capacity has dropped from 5600 to 3200. Production would fall behind possibly impacting conferences and deadlines. High number of jobs in queue (70K).

CMS – running without all available worker nodes.

Temporary Workaround

At ~5:25P on Tuesday June, 7 service was restored to an 'acceptable' level. Racks were turned on in a pre-determined order to reach 60% power capacity.

Room GCC 2 B operated at ~60% power capacity until 6:10A on Thursday, June 9.

Timeline of Events		
Date	**Time**	**Event**
6/7	3:13P	Received email from Vicky White that Adam Walters was reporting GCC B was experiencing cooling troubles. Group to consider graceful shutdown prior to shutting off power to the room. Email is supporting document.
6/7	3:18P	Jon Bakken emailed that CMS will look in to shutting down machines. Email is supporting Document.

6/7	3:20P	Adam Walters emailed “There is a cooling problem at GCC CRB and we are trying to stabilize the situation. We have shutdown 7 racks of computing. We may need to shutdown additional computing. The temperatures outside are contributing to the problem.” Email is supporting document.
6/7	3:32P	Adam Walters emailed “Computer Room B computing has been turned off. All cooling was lost. All other rooms at GCC are stable.” Email is supporting document.
6/7	3:33P	I responded in email that FCC1 is being established as a Critical Command Center. I also provided information for an audio bridge. Email is supporting document.
6/7	3:35P	Critical Incident Room established. Begin review of issue and impact. CMS noted that they were able to shut down 300 nodes ahead of the room power being shut off. Communication channels were established – I would be communicating to CD via email as well as all followers on Twitter. Margaret Votava would be communicating to experiment liaisons with the assumption that liaisons would communicate to experiments.
6/7	3:38P	Posted in SD-Broadcast: (6/7/2011 3:38:30 PM) **brianmck:** Cooling Emergency at GCC B - Computer Room B computing has been turned off. All cooling was lost. All other rooms at GCC are stable. (6/7/2011 3:38:52 PM) **brianmck:** FCC1 has been established as the command center (6/7/2011 3:39:13 PM) **brianmck:** or 1-866-740-1260 (6/7/2011 3:48:15 PM) **schmidt:** use code 1113472
6/7	3:40	Status report from Facility Operations – they were working to stabilize cooling. The heat/humidity of the room overwhelmed the CRAC cooling units. The CRAC units failed, resulting in power to the room being turned off. Several questions were raised (Was this a design problem, failure, other?). Adam Walters also reported that the concrete pads were being sprayed down with water. There were no alerts from the AC units.
6/7	3:45P – 4:05P	While to room was getting back to an acceptable temperature, the group began talking about power-up sequences, including what should come up first and what, if anything, wouldn’t come up at all today. Facilities was able to provide hard copies of the room which show racks and their locations. Due to power to the room being off, we were unable to remotely monitor power or temperature. During discussion it was noted by Anna Jordan there were some private networks in the GCC B. One was determined to be FeriCloud, but owners of the others were unknown.
6/7	3:52P	Email sent from Service Desk/tweet: All cooling has been lost at GCC Computer Room B. Computing has been turned off. All other rooms at GCC are stable. Updates to follow.
6/7	4:05P	GCC B room temperature 76 degrees.
6/7	4:24P	Posted in SD-Broadcast: (6/7/2011 4:24:50 PM) **Gerald Guglielmo:** anna is representing networking on the phone bridge and is muted, just give her a shout if you need something from networking.

6/7	4:33P	Jack Schmidt sent update to Senior Management: The room is now cooling down. Adam believes we can begin powering up systems. Discussing which racks to power up because we will not be able to power up all of it Current plan is to power up FermiGRID first. (Significant impact to FermiGRID jobs) Enabling power to monitoring equipment.
6/7	4:05P – 4:40P	Keith offered to leave FermiCloud down over night as it was a pilot. Racks 3027 – 3021 were noted as the hottest at the moment. Adam Walters believes we can turn up to on between 50% and 70% power capacity and there would be no issues overnight. The plan is to turn off individual breakers and turn on racks in the desired order. Discussions continued to determine sequence for power-up. The following order was determined: 1. Rack 3056 FermiGrid / System Monitoring 2. Racks 3061 and 3060 FNPC (check in with room) 3. Racks 3083; 3082; 3081; 3080; 3079; 3078; 3077; 3069; 3068; 3067; 3066; 3065; 3065; 3064; 3063; 3053; 3052; 3051; 3050; 3049; 3048; 3047; 3046; 3045; 3033; 3032; 3031 CMS Racks 3074; 3073; 3043; 3036; 3035 CDF Racks 3072; 3071; 3062; 3055; 3054; 3045; 3042; 3038; 3037; 3030 FNPC Racks 3059; 3058; 3057; 3056; 3041; 3040; 3039; 3034; 3029; 3028 D0 This plan is was determined to be the best way forward for overnight stability. It was also established that the following racks would remain powered down overnight – 3076, 3075; 3013, 3012, 3011, 3009, 3008, 3005, 3004, 3003, 3002, 3001, 3000, 3027, 3026, 3025, 3024, 3023,3022, 3021, 3020, 3019, 3018, 3017, 3016, 3015 and 3014. These racks are a mix of CDF, CMS, and D0 nodes. Also, FermiCloud would remain down overnight.
6/7	4:40P	Adam Walters left for GCC to perform the power-up sequence as noted above.

6/7	4:56P	I sent the following email to Senior Management: Adam and Tim are currently heading out to GCC. Trying to get to an estimate of 50% of power. All of the individual breakers will be turned off so the power-up sequence can be controlled. The current plan is to power up the following: Temperature Monitoring Racks 3056 – FermiGrid 3061 – FNPC 3060 – FNPC Keith volunteered to leave FermiCloud down overnight. There are no 24x7 services here. Also the following Racks won't be turned on tonight 3055/3054 (FNPC) and 3076/3075 (FCL). More to follow.
6/7	4:57P	Jason Allen reported that nodes were coming back up.
6/7	5:00P	Jack Schmidt sent an email to commence center participants: Hi! I am following up on the conversation held in the Critical incident Room. Please document the impact to your service from having GCC Room B trip. Please also document the impact to your service from the partial restoration. If possible reply to this message tonight- latest tomorrow morning by 9am.
6/7	5:03P	Adam Walters reported back to the room that racks 3056, 3061. 3062 and System Monitoring were restored. Restoration steps 1 and 2 were complete. 'Green light' to begin step 3 was given.
6/7	5:03P	Posted in SD-Broadcast: (6/7/2011 5:03:31 PM) **schmidt:** Turning systems on and watching temperatures
6/7	5:05P	FermiGrid service restored.
6/7	5:10P	FNPC service restored.
6/7	5:23P	All CMS nodes restored. CDF has 2250 slots available.
6/7	5:25P	Everything that was planned to be restored has been restored. Confirmed by the room.
6/7	5:30P	Adam reported to the room that there was a cooling imbalance and wanted to power up 2 racks on the north end of the room. At this time 3007 CMS and 3006 CDF were powered up.
6/7	5:42P – 6:00P	Adam arrived back at room and general discussion started taking place. Jason Allen and Keith Chadwick were investigating an failed Grid PDU. I confirmed with them that this component was not critical. Keith said he would address the issue tomorrow.
6/7	6:06P	David Fagan IMd Jon Bakken that the room appeared to be getting hot again. Adam Walters to drive to GCC to investigate. He would report any issues back to the room. It turned out that there were no issues and temperature was holding. General discussion followed which included solidifying plans for the following day.

6/7	6:06P	I sent the following email to Senior Management: Everything that has been planned to be turned on – has been turned on. GCC B is running at ~60% power capacity. Basically, 4 of the 6 rows have been turned on. The older nodes remain off. The following racks have <u>not</u> been turned on: 3013, 3012, 3011, 3009, 3008, 3005, 3004, 3003, 3002, 3001, 3000, 3027, 3026, 3025, 3024, 3023,3022, 3021, 3020, 3019, 3018, 3017, 3016, 3015 and 3014. These racks are a mix of CDF, CMS, and D0 nodes. The plan is to run in this capacity overnight (this is the level the group was comfortable with and would keep the experiments functioning at an 'acceptable' level). Facilities is set to receive any alerts should anything happen overnight. The service providers have been validating services and the only known issue is a PDU down in a FermiGrid rack. Keith was going to handle this tomorrow. We are re-grouping at 7:00A with Facilities to review data from overnight and to solidify the plan for the day.
6/7	6:08	Posted in SD-Broadcast: (6/7/2011 6:08:20 PM) **brianmck:** All racks have been turned on that we plan to be turned on. GCC B is running at ~60% power. Status will be evaluated in the AM.
6/7	6:31	Email sent from Service Desk/Tweet: Earlier today, there was an issue with cooling at GCC B. At this time, the temperature has stabilized and partial service has been restored. More information will be available tomorrow.
6/7	6:45P	Command Center Closed

6/8	7:00A – 8:00A	ITSM met with Facilities to get current status and to establish plan for the day. There were no issues overnight. It was agreed to that the plan was to prepare to shed power capacity. Facilities would continue to monitor and cool the pad with water. The following scenarios were established: **Scenario A** – Lose CRAC in SE corner after 1P Alternate 3 of 6 CMS racks in 3083-3077 **Scenario B** – Lose CRAC in SE corner before Noon Shut Down all of 3083 – 3077 **Scenario C** – Lose CRAC anywhere else Assess situation Also, racks 3011 CMS and 3002 CDF would be turned on to assist with keeping the CRAC units operating at a more balanced level.
6/8	8:00A-9:00A	The above plan was shared with Stakeholders. There was agreement on the approach. Scenario B was worked through with CMS to determine the which CMS racks would remain up. Patty also mentioned that the division spent a lot of money to avoid this type of situation and raised concerns about the rest of the summer. Margaret mentioned more racks were scheduled to be installed this fall. Qizhong mentioned that CDF want racks turned on tonight if possible versus tomorrow. There is a CDF conference on Monday. Jason Allen further elaborated on impact: There are 2 batch system for D0 currently impacted. One is running at 75% and the other is running at 90%. Adam Walters would send a status at noon and this group will meet again at 3:00P in FCC1 to determine when to power up the rest of the racks. Other impact included users finding and resubmitting failed jobs. Patty confirmed to operate in this state is more desirable than hard stops. There is also no risk of losing data, but production would fall behind possibly impacting conferences and deadlines. Margaret reported CDF had 70K jobs in queue.
6/8	8:39	Posted in SD Broadcast: (6/8/2011 8:39:57 AM) **brianmck:** GCC B will continue to operate under reduced capacity. The room will be monitored throughout the day. Determination will be made by EOD on restoring service to normal capacity.
6/8	8:45	Email sent from Service Desk/Tweet: GCC B will continue to operate under reduced capacity. The room will be monitored throughout the day. Determination will be made by EOD on restoring service to normal capacity.
6/8	8:50A	Adam Walters sent me an email confirming rack 3011 and 3002 have been restored.

6/8	9:38A	I sent the following email to Senior Management: Good morning, This morning the team met to understand the current situation and well as level-set on the plan for today. During the meeting, three scenarios were outlined: • Scenario A – Lose CRAC in SE corner after 1P ○ Action: Alternate shut down 3 of 6 CMS racks in 3083-3077 • Scenario B – Lose CRAC in SE corner before Noon ○ Action: Shut Down all of 3083 – 3077 (these are all CMS racks) • Scenario C – Lose CRAC anywhere else ○ Action: Assess situation Also during the meeting, it was determined that turning on 2 additional racks on the north end of the room would assist overall cooling by keeping the CRAC units on the north end running and balancing the room. The two racks that were turned on are 3011 (CMS) and 3002 (CDF). At noon today, Facilities will provide a status. At 3:00P today, the team will gather in FCC1 for status and to determine next steps. The room is operating at ~60 power capacity. The following racks remain powered off: 3013, 3012, 3009, 3008, 3005, 3004, 3003, 3001, 3000, 3027, 3026, 3025, 3024, 3023,3022, 3021, 3020, 3019, 3018, 3017, 3016, 3015 and 3014. These racks are a mix of CDF, CMS, and D0 nodes.
6/8	11:28A	Adam Walters emailed: Computer Room B is running fine at this point. No problems. Any change from turning on the 2 racks this morning is in the monitoring noise, which is not to say there was no benefit. The nice breeze and partial cloudiness is favorable.

6/8	11:43A	I sent the following email to Senior Management: Update from Adam Walters: Computer Room B is running fine at this point. No problems. Any change from turning on the 2 racks this morning is in the monitoring noise, which is not to say there was no benefit. The nice breeze and partial cloudiness is favorable.
6/8	11:44A	Posted in SD-Broadcast: (11:44:03 AM) **brianmck:** Update from Adam Walters re: GCC B - Computer Room B is running fine at this point. No problems.
6/8	11:53A	Email sent from Service Desk/Tweet: GCC B continues to run at 60% capacity. No issues have been detected. The room will continue to be monitored.
6/8	3:00P	Meeting for status in FCC1.

6/8	3:00P-3:45P	Status from Adam Walters: No issues with GCC B. The question was asked: When we would be comfortable turning things back on? Facilities - Tomorrow morning. Recommending not to do anything. Seen a capacity increase on the units. 3 units are currently running at 100%. Patty – bringing up another row for CDF or D0 would be helpful. Tim C – Temps are in the 70s until 4AM. They cross 80 at 9:00 tonight. They cross 90 at 5:00. Currently 90 degree…potentially up to 93. Jon Bakken – Is something still broken? Adam thinks it is a design issue, but its speculation. Possibly condensers being in a valley. Mark K – Is there a way to better understand what is a bad sign on the sensor? Keith – Return refrigerant temperature. Jason Allen – Have any steps been taken to determine root cause? No. Adam – We encountered these issues last summer when the temp was above 90 and sunny. Not able to determine wind affect. Facilities is getting a weather station for the pad. The CRACS started going 'hi head' and started shutting down. This was also observed last year. Refrigerant coming back was so hot that the compressor over heats. We shed 3 racks yesterday in FCC B last year. Patty – below 90 and not sunny = ok. Do we want people to stay in late or come in early to do this? The experiments biggest fear is for us to drop the power again. Is the issue only dependent on outside temp? Adam - Not sure. The room will still be heating up for the next couple hours, recommend waiting for the AM to turn things back on. Tim C: What time? Patty - 6AM. Keith C recommended to put thermocouples on the CRAC units to gather more data points. General discussion about short term workarounds: fans or dry ice? Tim C : Need short and long term plans to address issue. Wrap up: We will meeting in FCC1 at 6AM.
6/8	3:45P	Email sent from ServiceDesk/Tweet: GCC B will operate at 60% capacity overnight. The remaining systems will be powered up at 6:00AM on Thursday, June 9. Also Tweeted.

6/8	3:48P	I sent the following email to Senior Management: GCC B will operate at 60% capacity overnight. The plan is to be conservative for tonight. It is believed that the experiments would rather operate 'as-is' than risk a complete outage. The remaining systems will be powered up at 6:00AM on Thursday, June 9. The ITSM team will be in FCC1.
6/9	6:00A	Meeting room established / audio bridge opened
6/9	6:00A	Facilities restored power to remaining systems
6/9	6:00A	Jon Bakken Confirms power to CMS nodes
6/9	6:25A	Adam joins the meeting at FCC1. Reports no issues.
6/9	6:30A	Email sent from Service Desk/Tweet: All systems in GCC B have been powered on. The room is running at normal capacity. Thank you for your patience.
6/9	6:58A	I sent the following email to Senior Management: All systems were powered on at 6:00A as planned. The room is running at normal capacity.

Analysis Questions and Research
Who initially reported the issue?
Adam Walters
To who was the incident reported?
Victoria White
Was this a result or indirect result of a change request? If yes, what was the change request number?
No.
Outstanding Issues:
CMS: From Jon B - 13 nodes were broken (11 disk, 1 power supply, 1 memory) as a result of the high temperature, power down. There are about 1200 nodes total, so we are looking at about a 1% failure rate.

Incident Resolution

Resolved Via: Workaround.
An acceptable level of capacity was determined to restore service. The plan was to restore full capacity dependent on the weather. This plan was executed with no issues.